Available online at www.sciencedirect.com
ScienceDirect
Computers in Industry 59 (2008) 3–16
COMPUTERS IN INDUSTRY
www.elsevier.com/locate/compind

Review article

A review of structured document retrieval (SDR) technology to improve information access performance in engineering document management

S. Liu [a,*], C.A. McMahon [b], S.J. Culley [b]

[a] University of Strathclyde, CAD Centre, DMEM Department, 75 Montrose Street, Glasgow G1 1XJ, UK
[b] Department of Mechanical Engineering, University of Bath, Bath BA2 7AY, UK

Received 9 November 2006; received in revised form 27 June 2007; accepted 24 August 2007
Available online 24 October 2007

Abstract

Information retrieval (IR) is a well-established research and development area. Document formats such as SGML (Standard Generalised Mark-up Language) and XML (eXtensible Mark-up Language) have become widely used in recent years. Traditional IR systems demonstrate limitations when dealing with such documents, which motivated the emergence of structured document retrieval (SDR) technology intending to overcome these limitations. This paper reviews the work carried out from the inception to the development and application of SDR in engineering document management. The key issues of SDR are discussed and the state of the art of SDR to improve information access performance has been surveyed. A comparison of selected papers is provided and possible future research directions identified. The paper concludes with the expectation that SDR will make a positive impact on the process of engineering document management from document construction to its delivery in the future, and undoubtedly provide better information retrieval performance in terms of both precision and functionality.

Keywords: Literature review; Structured document retrieval; Information access performance; Engineering document management

Contents

* Corresponding author. Tel.: +44 141 548 2374; fax: +44 141 552 7986.
E-mail address: shaofeng.liu@strath.ac.uk (S. Liu).

0166-3615/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.compind.2007.08.001

1. Introduction

Engineering processes are highly creative and knowledge-intensive and comprise activities such as design, engineering analysis, manufacture and performance evaluation. As the complexity of these activities increases, so does the complexity of the information exchange and communication between engineers [1]. Several studies show that engineers spend as much as two-thirds of their time communicating in order to get input to their work and to output results from their work [2], and one-third of their time on searching for and accessing design information [3]. Many engineers seek information from plural sources such as documents, people and agents, depending on the stage of engineering process. For example, in the early design stage (such as conceptual design) when product features are vague, design engineers may focus more on seeking information from colleagues, communities and their own memories [4]. At the later stages such as detailed design and engineering analysis stages, engineers will rely more on formally recorded information such as reports, drawings, models and manuals [5,6].

It is believed that only 20% of formal information can be extracted from data warehouses comprising numeric data only, the other 80% of information is hidden in documents [7]. A similar observation has been made by Feldman [8], who claims that 80% of explicit knowledge in an enterprise can be found in documents. Therefore, document management has been recognised as a key topic in the information and knowledge management [9], which has been a well-established research field and has been successful in application for many areas. This paper provides a review on improving *document* access performance by integrating key approaches such as structure study and mark-up into an engineering domain. Readers who are interested in information seeking from *people* can refer to [10–12], study of how engineers' information seeking practices intertwine looking for informing documents with looking for informed people can be found in [2].

It is important to help engineers find the right information from documents at the right time in a way that most suits the engineers' profiles [13]. However, the information access performance of many document retrieval systems is limited by the technologies employed in the systems [14]. In traditional IR, systems make use of content information and focus on finding and delivering relevant whole documents (e.g. an entire paper or a Web page) in response to the user's needs. However, research has shown that the information needs of engineers are often not best met by the return of whole documents, but rather of the most relevant parts (of varying extent) of documents [15–17]. In particular, searching for information with traditional IR in very large documents is problematic, as is finding and collating specific information from multiple document sources. To remedy this, structured document retrieval (SDR) technology has established itself in recent years as an active field of research and development which is distinguished from traditional IR in making use of both structural and content information, therefore allowing users to retrieve the most relevant *components* of documents, i.e. document content that is more focused on the users' information needs, for example, a section of a book instead of an entire book [18–20]. Moreover, appropriate representation of a structured document allows for the retrieval process to return *aggregated* components, for example a set of sections, or all sections of the documents that are relevant to a query, instead of delivering the whole document [21]. Therefore, by taking advantage of the structural information, SDR has two major gains: increased functionality and increased precision [22–24]. SDR is becoming increasingly important to engineering information and knowledge management as structured document formats such as SGML (Standard Generalised Mark-up Language) and XML (eXtensible Mark-up Language) are becoming widely used in enterprises for document publishing and distribution, using Web and other technologies [25].

This paper provides a review on various research activities carried out in the past decade involving the SDR in engineering information management. The core of the paper identifies two key SDR academic research areas: document structure study and mark-up. A comparison of selected papers is carried out based on exploitation of structural information and mark-up for improvement of information access performance in terms of classification, indexing, querying and ranking, and presentation. The final part provides a discussion on the gaps in the research and suggests future directions for SDR research in engineering document management.

2. SDR key issues

This section discusses two key issues of SDR that have a strong impact on information access performance. One is document structure study. The other is document mark-up.

2.1. Document structure study

Document structure is about how content objects are organized in a document such as books, scientific articles and technical manuals [26]. The structure of documents provides a new source of information in addition to document content that IR systems may exploit to improve their search effectiveness. This section starts with the definition of document structure, and then reviews the work on analysis of document structure.